Estimating the rate of intersubtype recombination in early HIV-1 group M strains

Journal of Virology

Estimating the Rate of Intersubtype Recombination in Early HIV-1 Group M Strains

Melissa J. Ward, Samantha J. Lycett, Marcia L. Kalish, Andrew Rambaut and Andrew J. Leigh Brown
***J. Virol.* 2013, 87(4):1967. DOI: 10.1128/JVI.02478-12.**
Published Ahead of Print 12 December 2012.

JVI Journals.ASM.org

Estimating the Rate of Intersubtype Recombination in Early HIV-1 Group M Strains

Melissa J. Ward,[a] Samantha J. Lycett,[a] Marcia L. Kalish,[b] Andrew Rambaut,[a,c] Andrew J. Leigh Brown[a]

University of Edinburgh, Institute of Evolutionary Biology, Ashworth Laboratories, Edinburgh, United Kingdom[a]; Vanderbilt University, Vanderbilt Institute for Global Health, Nashville, Tennessee, USA[b]; Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA[c]

West Central Africa has been implicated as the epicenter of the HIV-1 epidemic, and almost all group M subtypes can be found there. Previous analysis of early HIV-1 group M sequences from Kinshasa in the Democratic Republic of Congo, formerly Zaire, revealed that isolates from a number of individuals fall in different positions in phylogenetic trees constructed from sequences from opposite ends of the genome as a result of recombination between viruses of different subtypes. Here, we use discrete ancestral trait mapping to develop a procedure for quantifying HIV-1 group M intersubtype recombination across phylogenies, using individuals' *gag* (p17) and *env* (gp41) subtypes. The method was applied to previously described HIV-1 group M sequences from samples obtained in Kinshasa early in the global radiation of HIV. Nine different p17 and gp41 intersubtype recombinant combinations were present in the data set. The mean number of excess ancestral subtype transitions (NEST) required to map individuals' p17 subtypes onto the gp14 phylogeny samples, compared to the number required to map them onto the p17 phylogenies, and vice versa, indicated that excess subtype transitions occurred at a rate of approximately 7×10^{-3} to 8×10^{-3} per lineage per year as a result of intersubtype recombination. Our results imply that intersubtype recombination may have occurred in approximately 20% of lineages evolving over a period of 30 years and confirm intersubtype recombination as a substantial force in generating HIV-1 group M diversity.

Strains of human immunodeficiency virus type 1 (HIV-1) separate into the following four phylogenetically distinct groups: M, N, and O, believed to have arisen via independent transmissions of simian immunodeficiency virus (SIV) from chimpanzees (1, 2), and group P, which is closely related to strains from gorillas (3). HIV-1 group M is the most common globally, being responsible for over 95% of infections worldwide. Within HIV-1 group M, several phylogenetically distinct subtypes (A, B, C, D, F, G, H, J, and K) (4), which are very divergent genetically, exist (5). Kinshasa, the capital of the Democratic Republic of the Congo (DRC), formerly Zaire, has been implicated as the epicenter of the HIV-1 group M epidemic (reviewed in reference 6). It is thought that the individual subtypes found in the city of Kinshasa emerged as a result of founder effects and geographical isolation (7–9).

HIV is a rapidly evolving pathogen with a high mutation rate resulting from an error-prone replication cycle (10, 11) and short generation time (12). HIV-1 virions contain two copies of the single positive strand of RNA which encodes the HIV-1 genome. Viruses which are identifiable as recombinants can arise through template switching during reverse transcription, when more than one genetically distinct virus is harbored in an infected cell at the same time (13). *In vitro* studies have estimated a minimum of 2.8 crossover events per genome per round of replication—an order of magnitude higher than the mutation rate (14). Mutation and recombination can thus both play a significant role in generating intrahost diversity in HIV-1 (5, 15).

Recombination in HIV-1 can occur between viruses of the same subtype (16, 17) and between viruses of different subtypes (18). Multiple infections of individuals with viruses of different subtypes, a prerequisite for intersubtype recombination, may result from a single transmission of genetically different viruses or a subsequent HIV infection acquired by an already infected individual. Intersubtype recombination was identified as a major mechanism for the generation of HIV-1 group M diversity by Robertson et al. (18), who reported numerous individuals from whom sequences from the *gag* and *env* regions (i.e., from opposite ends of the genome) were of different subtypes based upon phylogenetic analysis. Intersubtype recombinant viruses with the same breakpoints which are known to have caused infection in three or more epidemiologically unlinked individuals are known as circulating recombinant forms (CRFs) (4). At least 50 CRFs have now been characterized (see www.hiv.lanl.gov), and intersubtype recombinant viruses are thought to account for more than 20% of HIV cases worldwide (19).

Vidal et al. (7) analyzed HIV-1 group M sequences sampled in the DRC in 1997, finding discordant *gag* (p24) and *env* (V3-V5) subtypes in 29% of samples and providing evidence that most subtypes were involved in intersubtype recombination. Subsequently, Kalish et al. (9) found that over 25% of sampled infected hospital workers in Kinshasa in the mid-1980s had discordant *gag* (p17) and *env* (gp41) subtypes. However, while the prevalence of intersubtype recombinant HIV viruses can be estimated from sequencing studies, little is known about the frequency with which such viruses arise *in vivo* (20) and no estimates are available for the rate at which they contribute to HIV diversity at the interhost (population) level or how intersubtype recombination has operated historically.

Received 12 September 2012 **Accepted** 6 December 2012
Published ahead of print 12 December 2012
Address correspondence to Andrew J. Leigh Brown, A.Leigh-Brown@ed.ac.uk.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.02478-12.

doi:10.1128/JVI.02478-12
The authors have paid a fee to allow immediate free access to this article.

Previous phylogenetic approaches for investigating recombination have focused on detecting phylogenetic incongruences (e.g., see reference 21) rather than quantifying the amount of recombination between phylogenies or estimating recombination rates. In contrast, in this study, we used Bayesian discrete ancestral trait-mapping methods (22–26) to quantify intersubtype recombination in population-level phylogenies by comparing trees for protein-coding sequences at different ends of the HIV-1 genome. We calculated the number of excess ancestral subtype transitions required to map individuals' viral subtypes for sequences from one end of the genome onto the tree for the opposite end of the genome compared to the number required to map subtypes onto the tree for the correct end of the genome. This quantity was then scaled to estimate the rate (per year) at which lineages would be expected to undergo intersubtype recombination. The method was applied to the data set of Kalish et al. (9), comprising p17 *gag* and gp41 *env* sequences isolated from hospital workers in Kinshasa between 1984 and 1986. Due to the cocirculation of all subtypes (except B) in Kinshasa, this data set provided an unparalleled opportunity to investigate intersubtype recombination within a freely mixing population and allowed us to obtain a phylogenetic quantification of population-level HIV-1 group M intersubtype recombination.

MATERIALS AND METHODS

HIV-1 sequence data. All HIV-1 sequences for the *gag* p17 and *env* gp41 regions published by Kalish et al. (9) were downloaded from GenBank. The sequences had been obtained from hospital workers in Kinshasa between 1984 and 1986 as part of the Projet SIDA surveillance program by consensus sequencing of PCR-amplified RNA directly from serum samples. Data sets were created containing only sequences from persons (a total of 57) for whom both gp41 and p17 sequences were available (see Table S1 in the supplemental material). Sequences were aligned manually using BioEdit (27), and the alignments were 429 bp and 369 bp long for p17 and gp41, respectively. In order to assign subtypes to the p17 and gp41 sequences, reference sequences from the Los Alamos HIV Database (www.hiv.lanl.gov) were downloaded for each subtype as well as for CRFs 01 and 02. Chimpanzee (CPZ.CM.1998.CAM3.AF115393) gp41 and p17 sequences, which fall basal to HIV groups M and N (28), and reference sequences for group N and group O viruses (Ref.N.CM.95.YBF30.AJ006022 and Ref.O.BE.87.ANT70.L20587, respectively) were downloaded for use as outgroups.

Subtyping and preliminary phylogenetic analysis. Phylogenetic analyses were conducted to assign subtypes to the p17 and gp41 sequences, since some uncertainty had been reported in the neighbor-joining analysis of Kalish et al. (9). Maximum likelihood (ML) phylogenetic trees were constructed in PhyML (29) with 1,000 bootstrap replicates. A general time-reversible model (30) of nucleotide substitution was implemented, with gamma-distributed rate heterogeneity across sites and four rate categories. The effect of using different outgroups (group N or O or chimpanzee sequences) or a midpoint rooted tree was considered. Sequences were classified as a particular subtype if they belonged to a clade containing a reference sequence of that subtype and no reference sequences of any other subtype. Sequences which were basal to clades containing two or more subtypes were labeled as "unclassified."

Preliminary Bayesian phylogenetic analysis was carried out using Bayesian evolutionary analysis by sampling trees (BEAST) (31) to confirm the subtyping of sequences from the ML analysis and determine the molecular clock model providing the best fit to the data. A relaxed demographic prior (Bayesian skyline with 5 bins) (32) was implemented, and the SRD06 nucleotide substitution model (33) was used. Since precise sample date information was not available for the sequences, the mean substitution rate for the uncorrelated lognormal relaxed clock model was fixed to 1, returning branch lengths in units of substitutions per site. Markov chain Monte Carlo (MCMC) sampling took place every 10,000 generations over a period of 100 million generations in all BEAST runs, with a burn-in of 10 million generations. The chain traces were inspected in the Tracer software (31) (available from http://beast.bio.ed.ac.uk/Tracer) to indicate whether stationarity had been achieved, and multiple runs were compared for all analyses. Effective sample sizes (ESS) were greater than 200 for all parameters estimated.

Within-gene recombination analysis. In order to investigate whether recombination had taken place within the gene fragments encoding the gp41 and p17 proteins, a single breakpoint analysis (34, 35) was performed on the individual p17 and gp41 alignments using HyPhy (36) on the DataMonkey web server (37) (www.datamonkey.org).

Intersubtype recombination analysis. Sequences in the p17 and gp41 alignments were labeled according to both the p17 and gp41 subtypes for that individual, and intersubtype recombinant viruses were identified by a discrepancy between p17 and gp41 subtypes. Discrete ancestral trait mapping in BEAST was used to infer ancestral subtypes along the posterior phylogeny samples. Starting with the subtypes at the tips of the tree, transitions between ancestral subtypes were modeled as an asymmetric continuous-time Markov process (22). In order to quantify the amount of intersubtype recombination between opposite ends of the genome, both the p17 and gp41 subtypes were independently mapped onto the p17 and gp41 phylogeny samples. The same MCMC settings were used as described for the preliminary BEAST analysis.

The number of ancestral subtype transitions along each BEAST phylogeny sample was counted using the "Markov jumps" method described by Minin and others (23–26) (see Fig. S1 in the supplemental material). The number of excess ancestral subtype transitions (NEST) required to map subtypes onto the phylogeny for the wrong gene (e.g., p17 subtypes onto the gp41 tree), compared to the number required to map subtypes onto the phylogeny for the correct gene (e.g., p17 subtypes on the p17 tree), was calculated for 9,000 randomly paired gp41 and p17 posterior phylogeny samples. Intervals of 95% highest posterior density (HPD) (the narrowest Bayesian credible intervals containing 95% of the data) were calculated for the NEST across the paired phylogeny samples.

In the absence of intersubtype recombination, the number of ancestral subtype transitions would be the same when subtypes for one end of the genome (e.g., p17) were mapped onto phylogenies for both the p17 and gp41 regions and the 95% HPD interval for the NEST would be centered on zero. Since intersubtype recombination events create excess ancestral subtype transitions when individuals' viral subtypes from one part of the genome are mapped onto the tree for another part of the genome, the NEST allows the amount of intersubtype recombination which can be detected between the two genome regions to be quantified. We rescaled the NEST to estimate the rate (per lineage per year) at which excess ancestral subtype transitions occurred as a result of intersubtype recombination. For each pair of phylogeny samples, the NEST was divided by the sum of the branch lengths (in units of substitutions per site) of the tree from the opposite end of the genome to the subtypes being mapped (e.g., gp41 tree when mapping p17 subtypes) and then multiplied by an estimate of the rate of HIV-1 nucleotide substitution of 2.47×10^{-3} substitutions/site/year (38) (see Fig. S2 in the supplemental material).

The potential for difficulty in assigning subtypes to have introduced error into the analysis was investigated by repeating the analysis using alternative subtype labeling for sequences which were difficult to classify as well by labeling sequences by the clade to which they belonged at a cutoff defined at the root of the subtype A clade in the maximum clade credibility (MCC) tree (see Fig. S3 and S4 in the supplemental material). Defining a cutoff near the root of the subtype A clade resulted in 10 clades being defined on the gp41 tree and 7 or 10 clades on the p17 tree, i.e., similar to the number of subtypes identified in the maximum likelihood analysis.

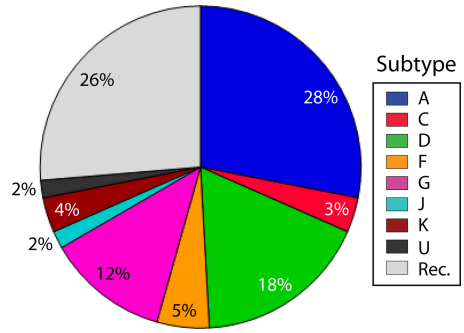

FIG 1 Subtype distribution of HIV-1 group M in Kinshasa. The gp41 and p17 regions of HIV-1 group M were sequenced for 57 individuals by Kalish et al. (9). Percentages of individuals infected with potentially pure viruses (i.e., with the same gp41 and p17 subtype) of a given subtype on the basis of maximum likelihood phylogenetic analysis are reported. A total of 26% of the viruses were classified as recombinant (Rec.) on the basis of different subtypes having been assigned to the p17 and gp41 regions.

RESULTS

Intersubtype recombinant viruses in Kinshasa. "Potentially pure" viruses, for which an individual's p17 and gp41 sequences were of the same subtype, were present in the Kinshasa data set for all subtypes except H, which is globally rare, and B, which does not appear among African sequences from this time (39). Discordant p17 and gp41 subtypes were found in 26% of individuals under the maximum likelihood phylogenetic analysis (Fig. 1), in line with a previous analysis of this data set (9). Nine different discordant p17 and gp41 subtype combinations were present in the data set (Table 1), and subtypes A, D, F, G, H, and J were involved in intersubtype recombinations. Subtype A, which was the most frequently occurring potentially pure subtype, was also the most commonly represented subtype among the recombinant viruses, with 11 out of the 15 (73%) recombinant viruses having a p17 or gp41 sequence of subtype A. The most frequently occurring recombinant virus was of type A_G (3 out of 15, i.e., 20% of the recombinant viruses). Additionally, three viruses were labeled as recombinants since their p17 sequences were of subtype A while their gp41 sequences formed a clade of their own, clustering in the maximum likelihood trees with reference sequences of type CRF 01.

Discrete ancestral trait mapping of p17 and gp41 subtypes was performed upon sets of BEAST phylogenies for the p17 and gp41 regions, and limited evidence for clustering of the recombinant viruses was observed in the maximum clade credibility (MCC) trees (Fig. 2). The two gp41 sequences from D_F viruses clustered together in the gp41 tree (posterior probability P value = 0.902), and their p17 sequences were also sister lineages ($P = 0.686$). In the p17 tree, two of the three A_CRF01 viruses clustered together ($P = 0.917$) and also clustered in the gp41 tree ($P = 0.998$). Since clustered HIV sequences are often considered to be epidemiologically linked (40), these clusters may represent a single intersubtype recombination event, followed by transmission of the recombinant virus. Clusters of intersubtype recombinant viruses arising from a single recombination event, followed by transmission of the recombinant virus, would incur only one additional ancestral subtype transition in our method for quantifying recombination, whereas multiple independent intersubtype recombinations across the tree would require further additional transitions.

Within-gene recombination analysis. No evidence of recombination was detected within the p17 or gp41 alignments using a single breakpoint analysis (35) under either the Bayesian informa-

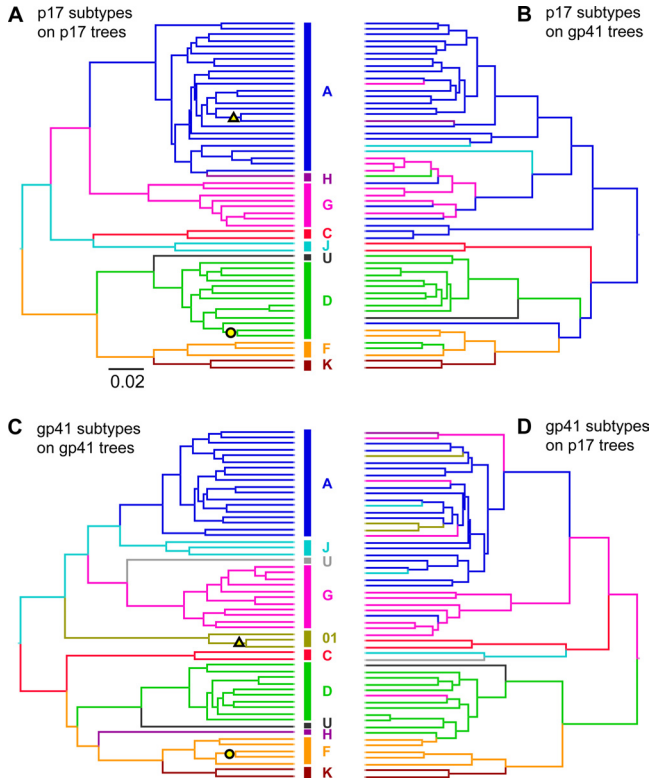

FIG 2 Maximum clade credibility (MCC) trees for the Kinshasa 1984 data set, colored by ancestral subtype. Maximum clade credibility trees were constructed using BEAST. Branches were colored according to inferred ancestral subtypes, mapping individuals p17 subtypes onto the p17 tree samples (A), patients' p17 subtypes onto the gp41 trees (B), gp41 subtypes onto the gp41 trees (C), and gp41 subtypes onto the p17 trees (D). The number of p17 and gp41 subtype transitions (Markov jumps) across the tree was recorded for each posterior phylogeny sample. Clustering of recombinant sequences can be observed in the MCC trees for two D_F individuals (marked with circles; posterior probabilities of being sister lineages are 0.686 and 0.902 in the p17 and gp41 trees, respectively) and two A_CRF01 individuals (marked with triangles; posterior probabilities of being sister lineages are 0.917 and 0.998 in the p17 and gp41 trees, respectively). Branch lengths are in units of substitutions per site.

TABLE 1 Frequency of recombinant types[a]

p17 subtype	gp41 subtype	Frequency
A	G	3
A	H	1
A	J	2
A	CRF01	3
D	F	2
D	G	1
G	A	1
H	A	1
J	U	1

[a] The p17 and gp41 subtypes of discordant sequences and their frequency of occurrence in the data set are reported. Note that "U" denotes an unclassifiable sequence and "CRF01" denotes the circulating recombinant form previously known as subtype E, which forms a distinct clade at the 3′ end of the genome.

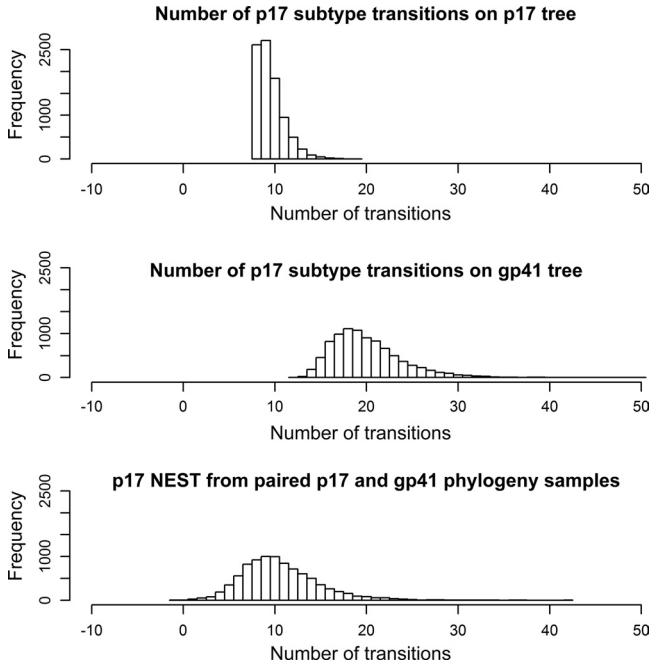

FIG 3 Number of inferred p17 ancestral subtype changes across phylogeny samples. The number of ancestral subtype transitions across phylogeny samples was inferred using ancestral trait mapping in BEAST, mapping p17 subtypes onto the p17 and gp41 phylogeny samples. The number of excess subtype transitions (NEST) required to map the ancestral p17 subtypes onto the phylogeny for the "wrong" gene, compared to the number required for mapping them onto the correct phylogeny, was calculated across paired phylogeny samples. Histograms indicate the numbers of subtype transitions across 9,000 post-burn-in samples of phylogenies.

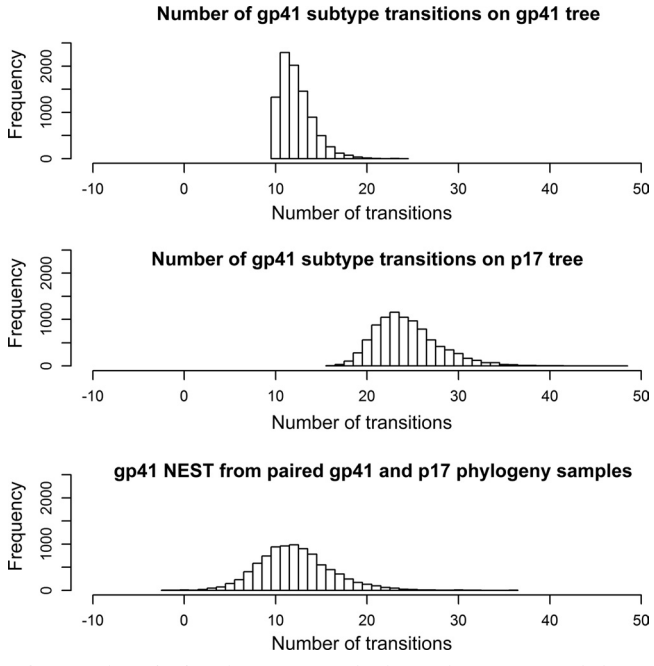

FIG 4 Number of inferred gp41 ancestral subtype changes across phylogeny samples. The number of ancestral subtype transitions across phylogeny samples was inferred using ancestral trait mapping in BEAST, mapping gp41 subtypes onto the p17 and gp41 phylogeny samples. The number of excess subtype transitions (NEST) required to map the ancestral gp41 subtypes onto the phylogeny for the "wrong" gene, compared to the number required for mapping them onto the correct phylogeny, was calculated across paired phylogeny samples. Histograms indicate the number of jumps across 9,000 post-burn-in samples of phylogenies.

tion criterion (BIC) or the corrected Akaike information criterion (AICc).

Quantifying intersubtype recombination. Discrete trait-mapping methods in BEAST have previously been used for phylogeographic analysis of viral sequence data (e.g., see references 22, 26, and 41) but can also be used to infer other ancestral character state data onto phylogenies, such as host species (42). Such methods have recently been applied to investigate reassortment in swine influenza (43). Here, we propose a new method for investigating intersubtype recombination in HIV using Markov jump counting of ancestral subtype transitions. If there are k different states at the tips of a phylogeny, the minimum number of ancestral state transitions observed across the phylogeny is $k - 1$. When HIV-1 group M subtypes were mapped onto the BEAST phylogeny for the correct gene (i.e., p17 subtypes on the p17 tree or gp41 subtypes on the gp41 tree), the number of ancestral subtype transitions across the tree lay toward this minimum number (Fig. 3 and 4; see also Fig. S1 in the supplemental material) and phylogenetic uncertainty accounted for instances in which more transitions were required. A greater number of ancestral subtype transitions were required to map individuals gp41 or p17 subtypes onto phylogeny samples for the other end of the genome (p17 subtypes on the gp41 tree or gp41 subtypes on the p17 tree) than to map them onto phylogeny samples for the correct gene (Fig. 3 and 4) as a result of intersubtype recombination. Intersubtype recombination may thus be detected from the phylogenies in this way.

The mean number of excess ancestral subtype transitions (NEST) required to map the p17 subtypes onto the gp41 phylogeny samples, compared to the number required to map the p17 subtypes onto the p17 phylogeny samples, was 10.55 (95% HPD interval, 2 to 18). The mean NEST for mapping gp41 subtypes onto p17 and gp41 phylogeny samples was 12.18 (95% HPD interval, 5 to 20). When the NEST was rescaled as described in Materials and Methods and Fig. S2 in the supplemental material, the rate at which excess ancestral substitutions arose was estimated to be 6.93×10^{-3} per lineage per year (95% HPD interval, 2.39×10^{-3} to 1.30×10^{-2}) using p17 subtype labels and 8.11×10^{-3} per lineage per year (95% HPD interval, 3.11×10^{-3} to 1.41×10^{-2}) using gp41 subtype labels.

The substantial overlap of the HPD intervals for NEST and for the rescaled NEST indicated that the estimates obtained using p17 and gp41 subtypes were not significantly different. Slight differences may have arisen because different numbers of gp41 subtypes and p17 subtypes (9 and 11, respectively) were present in the data set. HPD intervals were similar (see Table S2 in the supplemental material) when alternative subtype labelings were used for sequences which were difficult to classify and when clades were defined from a predetermined cutoff point along the tree (see Fig. S3 and S4 in the supplemental material), indicating that our results were robust to potential errors in subtyping of sequences.

DISCUSSION

Understanding recombination as an ancestral process is important for unraveling the evolutionary history of HIV and explaining the pattern of HIV diversity (44). In addition, recombination can

confound phylogenetic analyses which assume that a single evolutionary tree applies to the whole of an alignment (45), leading to false positives when detecting sites under positive selection (46, 47) and affecting estimates of divergence dates (48, 49), or at least increasing the variance of such estimates (50).

Although previous studies have investigated crossover rates *in vitro*, these do not measure the rate at which intersubtype recombination contributes to HIV diversity at the interhost phylogenetic level, a contribution which arises from a more complex, composite process. For an intersubtype recombination event to be detected from a population-level phylogeny, an individual must firstly be infected with viruses of more than one subtype, an intersubtype recombination must take place, and the resulting recombinant virus must be viable and become the dominant strain within an individual. While procedures have been developed for detecting recombination on the basis of phylogenetic discordance (e.g., see reference 51), methods for quantifying recombination across phylogenies are lacking. It has therefore been difficult to compare the rate of recombination and other evolutionary processes, such as nucleotide substitution, which contribute to the observed diversity of HIV-1 group M in populations such as Kinshasa.

We have been able to obtain an ancestral quantification of intersubtype recombination in HIV-1 group M. This extends earlier studies (7, 9), which simply reported the prevalence of viruses with discordant subtypes. By analyzing viral sequence data from Kinshasa, where almost all HIV-1 group M subtypes cocirculate, the observed recombination events can reasonably be assumed to have occurred within this population.

Our intersubtype recombination rate estimate of 6.93×10^{-3} to 8.11×10^{-3} excess ancestral subtype transitions per lineage per year was obtained by mapping individuals' subtypes from one end of the genome onto phylogenies for the other end of the genome. The rate estimate can be compared to other evolutionary processes, such as the HIV-1 nucleotide substitution rate, which has previously been estimated as 2.47×10^{-3} substitutions per site per year (38). The HIV-1 genome is 9,700 bp in length, and thus, nucleotide substitutions are expected to occur approximately 24 times ($9{,}700 \times 2.47 \times 10^{-3}$) per lineage per year across a single genome. Although our results indicate that the rate at which excess ancestral subtype transitions arise between different ends of the genome due to intersubtype recombination is considerably lower, intersubtype recombination has far greater potential for instantly generating highly novel HIV-1 group M virus strains than does the gradual accumulation of nucleotide substitutions and poses a significant problem for vaccine design (52). We can also use a Poisson process to calculate the probability that a lineage evolving for a given period of time would have undergone at least one intersubtype recombination event (e.g., $1 - \exp[-6.93 \times 10^{-3} \times \text{time period}]$ or $1 - \exp[-8.11 \times 10^{-3} \times \text{time period}]$) and estimate that 18.8 to 21.6% of lineages evolving for 30 years in the population studied would undergo intersubtype recombination.

The amount of HIV-1 group M intersubtype recombination observed in this study must be an underestimate of the actual amount within this population. The use of just two sections of the genome for identifying recombinant viruses means that isolates can be designated potentially pure by having the same gp41 and p17 subtype but, in fact, contain a section derived from a different parental subtype in another region. Furthermore, more than one crossover may occur along the HIV genome during reverse transcription and the method used here will not provide a measure of this. Recombination can also lead to viruses with sections derived from more than two parental subtypes, and such mosaic genomes would not be detected in this study. It must also be noted that the hospital workers studied may have been at a higher risk of multiple infection than the general Kinshasa population due to a lack of universal precautions to prevent blood-to-blood transmission through the course of their work. However, the HIV-1 group M prevalence estimate of 3.5% for hospital workers in Kinshasa in 1984 (9) is in line with estimates for childbearing women and blood donors (both 3.1%) in Kinshasa in 1997 (53). These findings suggested that seroprevalence of HIV-1 had stabilized in Kinshasa since the 1980s and that the hospital workers did not exhibit higher levels of HIV infection than the general population.

Experimental studies of intersubtype recombination in HIV-1 group M may provide a basis for further investigations using variants of our method. For example, Chin et al. (20) compared *in vitro* rates of intra- and intersubtype recombination among subtype B and C viruses and attributed the 10-fold-lower intersubtype recombination rate to a three-nucleotide difference in the dimerization initiation signal (DIS) region between the subtypes. It has been reported that HIV-1 group M subtypes fall into two groups with respect to DIS sequence, with B and D having the motif GCGCGC and A, C, F, G, H, and J possessing the motif GTGCAC (54, 55). Chin et al. (20) postulated that subtypes with different DIS motifs would recombine less frequently than subtypes which had the same motif. Although there were too few sequences of each subtype to test this hypothesis on the data set we analyzed, given additional data, our methods may be extended to compare rates of recombination between different HIV-1 group M subtypes.

Other studies (56–58) have shown that recombination breakpoints are not distributed randomly across the HIV-1 genome and have provided evidence for recombination "hot spots" as well as regions in which recombination is limited. Such patterns are thought to be due to a combination of mechanistic processes and selection. Given sequence data from multiple genomic regions, in the future, the NEST method may be used to quantify and compare the amounts of recombination between several different parts of the genome. These analyses may elucidate *in vivo* restrictions to recombination between different HIV-1 group M subtypes and be reconciled with the findings of studies of constraints on intersubtype recombination.

In conclusion, we have provided a phylogenetic quantification of ancestral HIV-1 group M intersubtype recombination in a population in which viruses of many subtypes are cocirculating. Our rate estimate yields predictions consistent with the observation that a substantial proportion of global HIV infections are caused by intersubtype recombinant viruses. In the future, biological questions, such as the level of multiple infection which would be required to observe a given intersubtype recombination rate for our data set, could be investigated. The NEST method may also be compared to population genetic measures of recombination or ancestral recombination graphs (ARGs) for samples of sequences (59). Although discrete trait-mapping methods for quantifying recombination lend themselves most naturally to viruses such as HIV and influenza, which can be classified into distinct subtypes (e.g., see reference 43), in principle, they can be applied to any

virus by defining clades along the phylogenies and labeling the tips according to the clade to which they belonged.

ACKNOWLEDGMENTS

We thank Emma Hodcroft for help with the initial sequence alignment.

M.J.W. performed the analysis and drafted the paper. M.L.K. obtained the sequences. S.J.L. and A.R. developed the methodology, and A.J.L.B. conceived the study and edited the manuscript. All authors have seen and approved the manuscript.

This work was supported by a BBSRC Doctoral Training award (to M.J.W.) and the Wellcome Trust (grant number 092807).

REFERENCES

1. **Gao F, Bailes E, Robertson DL, Chen YL, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH.** 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes.* Nature **397**:436–441.
2. **Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH.** 2001. The origins of acquired immune deficiency syndrome viruses: where and when? Philos. Trans. R. Soc. Lond. B Biol. Sci. **356**:867–876.
3. **Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, Damond F, Robertson DL, Simon F.** 2009. A new human immunodeficiency virus derived from gorillas. Nat. Med. **15**:871–872.
4. **Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B.** 2000. HIV-1 nomenclature proposal. Science **288**:55–57.
5. **Rambaut A, Posada D, Crandall KA, Holmes EC.** 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. **5**:52–61.
6. **Sharp PM, Hahn BH.** 2011. Origins of HIV and the AIDS pandemic. Cold Spring Harb. Perspect. Med. **1**:a006841. doi:10.1101/cshperspect.a006841.
7. **Vidal N, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, Delaporte E, Peeters M.** 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J. Virol. **74**:10498–10507.
8. **Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC.** 2001. Human immunodeficiency virus: phylogeny and the origin of HIV-1. Nature **410**:1047–1048.
9. **Kalish ML, Robbins KE, Pieniazek D, Schaefer A, Nzilambi N, Quinn TC, St Louis ME, Youngpairoj AS, Phillips J, Jaffe HW, Folks TM.** 2004. Recombinant viruses and early global HIV-1 epidemic. Emerg. Infect. Dis. **10**:1227–1234.
10. **Battula N, Loeb LA.** 1976. On the fidelity of DNA replication. Lack of exodeoxyribonuclease activity and error-correcting function in avian myeloblastosis virus DNA polymerase. J. Biol. Chem. **251**:982–986.
11. **Preston BD, Poiesz BJ, Loeb LA.** 1988. Fidelity of HIV-1 reverse transcriptase. Science **242**:1168–1171.
12. **Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD.** 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271**:1582–1586.
13. **Hu WS, Temin HM.** 1990. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc. Natl. Acad. Sci. U. S. A. **87**:1556–1560.
14. **Zhuang JL, Jetzt AE, Sun GL, Yu H, Klarmann G, Ron Y, Preston BD, Dougherty JP.** 2002. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J. Virol. **76**:11273–11282.
15. **Onafuwa-Nuga A, Telesnitsky A.** 2009. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. Microbiol. Mol. Biol. Rev. **73**:451–480.
16. **Liu SL, Mittler JE, Nickle DC, Mulvania TM, Shriner D, Rodrigo AG, Kosloff B, He X, Corey L, Mullins JI.** 2002. Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. J. Virol. **76**:10674–10684.
17. **Yang OO, Daar ES, Jamieson BD, Balamurugan A, Smith DM, Pitt JA, Petropoulos CJ, Richman DD, Little SJ, Leigh Brown AJ.** 2005. Human immunodeficiency virus type 1 clade B superinfection: evidence for differential immune containment of distinct clade B strains. J. Virol. **79**:860–868.
18. **Robertson DL, Sharp PM, McCutchan FE, Hahn BH.** 1995. Recombination in HIV-1. Nature **374**:124–126.
19. **Gao Y, Abreha M, Nelson KN, Baird H, Dudley DM, Abraha A, Arts EJ.** 2011. Enrichment of intersubtype HIV-1 recombinants in a dual infection system using HIV-1 strain-specific siRNAs. Retrovirology **8**:5.
20. **Chin MPS, Rhodes TD, Chen J, Fu W, Hu WS.** 2005. Identification of a major restriction in HIV-1 intersubtype recombination. Proc. Natl. Acad. Sci. U. S. A. **102**:9002–9007.
21. **Nagarajan N, Kingsford C.** 2011. GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res. **39**:e34.
22. **Lemey P, Rambaut A, Drummond AJ, Suchard MA.** 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. **5**:e1000520. doi:10.1371/journal.pcbi.1000520.
23. **Minin VN, Suchard MA.** 2008. Counting labeled transitions in continuous-time Markov models of evolution. J. Math. Biol. **56**:391–412.
24. **Minin VN, Suchard MA.** 2008. Fast, accurate and simulation-free stochastic mapping. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363**:3985–3995.
25. **O'Brien JD, Minin VN, Suchard MA.** 2009. Learning to count: robust estimates for labeled distances between molecular sequences. Mol. Biol. Evol. **26**:801–814.
26. **Talbi C, Holmes EC, De Benedictis P, Faye O, Nakoune E, Gamatie D, Diarra A, Elmamy BO, Sow A, Adjogoua EV, Sangare O, Dundon WG, Capua I, Sall AA, Bourhy H.** 2009. Evolutionary history and dynamics of dog rabies virus in western and central Africa. J. Gen. Virol. **90**:783–791.
27. **Hall TA.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41**:95–98.
28. **Hahn BH, Shaw GM, Cock KMD, Sharp PM.** 2000. AIDS as a zoonosis: scientific and public health implications. Science **287**:607–614.
29. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–321.
30. **Tavaré S.** 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. **17**:57–86.
31. **Drummond AJ, Rambaut A.** 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**:214.
32. **Drummond AJ, Rambaut A, Shapiro B, Pybus OG.** 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. **22**:1185–1192.
33. **Shapiro B, Rambaut A, Drummond AJ.** 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. **23**:7–9.
34. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW.** 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics **22**:3096–3098.
35. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW.** 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. **23**:1891–1901.
36. **Kosakovsky Pond SL, Frost SDW, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676–679.
37. **Delport W, Poon AFY, Frost SDW, Pond SLK.** 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics **26**:2455–2457.
38. **Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JMM, Kalengayi RM, Van Marck E, Gilbert MTP, Wolinsky SM.** 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature **455**:661–664.
39. **Vermund SH, Leigh Brown AJ.** 2012. The HIV epidemic: high-income countries, p 385–408. *In* Bushman FD, Nabel GJ, Swanstrom R (ed), Human immunodeficiency virus. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
40. **Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ.** 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. **5**:e50. doi:10.1371/journal.pmed.0050050.
41. **Raghwani J, Rambaut A, Holmes EC, Hang VT, Hien TT, Farrar J, Wills B, Lennon NJ, Birren BW, Henn MR, Simmons CP.** 2011. Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. PLoS Pathog. **7**:e1002064. doi:10.1371/journal.ppat.1002064.

42. **Weinert L, Welch JJ, Suchard M, Lemey P, Rambaut A, Fitzgerald JR.** 2012. Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an association with the spread of domestication. Biol. Lett. **8**:829–832.
43. **Lycett SJ, Baillie G, Coulter E, Bhatt S, Kellam P, McCauley JW, Wood JLN, Brown IH, Pybus OG, Leigh Brown AJ,** Combating Swine Influenza Initiative (COSI) Consortium. 2012. Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. J. Gen. Virol. **93**:2326–2336.
44. **Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme AM.** 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. **81**:8543–8551.
45. **Posada D, Crandall KA.** 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. **54**:396–402.
46. **Anisimova M, Nielsen R, Yang ZH.** 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229–1236.
47. **Shriner D, Nickle DC, Jensen MA, Mullins JI.** 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. **81**:115–121.
48. **Schierup M, Hein J.** 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics **156**:879–891.
49. **Worobey M.** 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. **18**:1425–1434.
50. **Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, Worobey M, Vandamme AM.** 2004. The molecular population genetics of HIV-1 group O. Genetics **167**:1059–1068.
51. **Posada D, Crandall KA.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U. S. A. **98**:13757–13762.
52. **Burke DS.** 1997. Recombination in HIV: an important viral evolutionary strategy. Emerg. Infect. Dis. **3**:253–259.
53. **Mulanga-Kabeya C, Nzilambi N, Edidi B, Minlangu M, Tshimpaka T, Kambembo L, Atibu L, Mama N, Ilunga W, Sema H, Tshimanga K, Bongo B, Peeters M, Delaporte E.** 1998. Evidence of stable HIV seroprevalences in selected populations in the Democratic Republic of the Congo. AIDS **12**:905–910.
54. **Andersen ES, Jeeninga RE, Damgaard CK, Berkhout B, Kjems J.** 2003. Dimerization and template switching in the 5′ untranslated region between various subtypes of human immunodeficiency virus type 1. J. Virol. **77**:3020–3030.
55. **Paillart JC, Dettenhofer M, Yu XF, Ehresmann C, Ehresmann B, Marquet R.** 2004. First snapshots of the HIV-1 RNA structure in infected cells and in virions. J. Biol. Chem. **279**:48397–48403.
56. **Archer J, Pinney JW, Fan J, Simon-Loriere E, Arts EJ, Negroni M, Robertson DL.** 2008. Identifying the important HIV-1 recombination breakpoints. PLoS Comput. Biol. **4**:e1000178. doi:10.1371/journal.pcbi.1000178.
57. **Simon-Loriere E, Galetto R, Hamoudi M, Archer J, Lefeuvre P, Martin DP, Robertson DL, Negroni M.** 2009. Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. PLoS Pathog. **5**:e1000418.
58. **Galli A, Kearney M, Nikolaitchik OA, Yu S, Chin MPS, Maldarelli F, Coffin JM, Pathak VK, Hu WS.** 2010. Patterns of human immunodeficiency virus type 1 recombination *ex vivo* provide evidence for coadaptation of distant sites, resulting in purifying selection for intersubtype recombinants during replication. J. Virol. **84**:7651–7661.
59. **Bloomquist EW, Suchard MA.** 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst. Biol. **59**:27–41.